

Exploiting Web Images for Dataset Construction: A Domain Robust Approach

Yazhou Yao, *Student Member, IEEE*, Jian Zhang, *Senior Member, IEEE*, Fumin Shen, *Member, IEEE*,
Xiansheng Hua, *Fellow, IEEE*, Jingsong Xu, and Zhenmin Tang

Abstract—Labelled image datasets have played a critical role in high-level image understanding; however the process of manual labelling is both time-consuming and labor intensive. To reduce the cost of manual labelling, there has been increased research interest in automatically constructing image datasets by exploiting web images. Datasets constructed by existing methods tend to have a weak domain adaptation ability, which is known as the “dataset bias problem”. To address this issue, we present a novel image dataset construction framework which can be generalized well to unseen target domains. In specific, the given queries are first expanded by searching in the Google Books Ngrams Corpus to obtain a richer semantic description, from which the visually non-salient and less relevant expansions are filtered out. By treating each unfiltered expansion as a “bag” and the retrieved images as “instances”, image selection can be formulated as a multi-instance learning problem with constrained positive bags. We propose to solve the employed problems by the cutting-plane and concave-convex procedure (CCCP) algorithm. Using this approach, images from different distributions will be retained while noisy images will be filtered out. To verify the effectiveness of our proposed approach, we build a domain-robust image dataset with 20 categories, which we refer to as DRID-20. We compare DRID-20 with three publicly available datasets STL-10, CIFAR-10 and ImageNet. The experimental results confirm the effectiveness of our dataset in terms of image classification ability, cross-dataset generalization ability and dataset diversity. We further run object detection on PASCAL VOC 2007 using our data, and the results demonstrate the superiority of our method to the weakly supervised and web-supervised state-of-the-art detection methods.

Index Terms—Domain robust, multiple query expansions, image dataset construction, MIL

I. INTRODUCTION

With the development of Internet, we have entered the era of big data. It is consequently a natural idea to leverage the large scale yet noisy data on the web for various vision tasks [1], [3], [4], [5]. Methods of exploiting web images for automatically image dataset construction have recently become a hot topic [12], [23], [28], [17] in the field of multimedia processing. Existing methods [12], [23], [17] usually use an iterative mechanism in the process of image selection, but these datasets tend to be statistically problematic because of the visual feature distribution of images selected in this way,



Fig. 1: Most discriminative images from four different datasets.

which is known as the dataset bias problem [21], [25], [41]. Fig. 1 shows the “airplane” images from four different image datasets. We observe some significant differences: PASCAL [10] shows “airplanes” from the flying viewpoint, while SUN [34] tends to show distant views at the airport; Caltech [32] has a strong preference for side views and ImageNet [2] is rich in diversity, but mainly contains close-range views. Classifiers learned from these datasets usually perform poorly in domain adaptation [21]. To address this problem, a large number of domain adaptation approaches that explicitly cope with the noisy labels of web images have been proposed for various vision tasks [26], [31]. The images are partitioned into a set of clusters; each cluster is treated as a “bag” and the images in each bag as “instances”. As a result, these tasks can be formulated as a multi-instance learning (MIL) problem, and different MIL methods have been proposed in [26], [31]. However, the yield for all of these methods is limited by the restriction of diversity which provided by image search engine with a single query.

To obtain high accuracy and diverse candidate images, as well as to overcome the download restrictions of the image search engine, [9], [28] proposed the use of multiple query expansions instead of a single query to collect candidate images from the image search engine. The issue remains that these methods still use iterative mechanisms in the process of

Y. Yao, J. Zhang and J. Xu are with the Global Big Data Technologies Center, University of Technology Sydney, NSW 2007, Australia.

F. Shen is an associate Professor in School of Computer Science and Engineering, University of Electronic Science and Technology of China.

X. Hua is a researcher/senior director in Alibaba Group, Hangzhou, China.

Z. Tang is a Professor in School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.

Manuscript received ; revised .

image selection, which leads to the dataset bias problem [21], [25], [41].

Motivated by the situation described above, we target the construction of an image dataset in a scalable way, while ensuring robustness and accuracy. The basic idea is to leverage multiple query expansions for the initial candidate image collection and to use MIL methods for selecting images from different distributions. To obtain multiple query expansions, we expand each query to a set of query expansions, which results in most of the noisy expansions being filtered out. After obtaining the raw image dataset with unfiltered query expansions, MIL methods are applied to filter individual and group noisy images. To verify the effectiveness of our proposed approach, we build an image dataset with 20 categories. We compare the image classification ability, cross-dataset generalization ability and dataset diversity of our dataset with three manually labelled image datasets, CIFAR-10, STL-10 and ImageNet, to demonstrate the domain robustness of our dataset. We also report the results of object detection on PASCAL VOC 2007, and then compare the object detection ability of our method with four baseline methods.

Our main contributions are summarized as follows:

[1.] To the best of our knowledge, we are the first to propose the automatic construction of a domain-robust image dataset. Our proposed approach, based on multiple query expansions and multi-instance learning, considers the source of candidate images and retains images from different distributions. The dataset constructed by our approach thus efficiently alleviates the dataset bias problem.

[2.] To suppress the search error and unfiltered noisy query expansions induced noisy images, we formulate image selection as multi-instance learning problems and propose to solve the associated optimization problems by the cutting-plane and concave-convex procedure (CCCP) algorithm, respectively.

[3.] We have released our image dataset DRID-20 on website: <https://drive.google.com/drive/folders/0B7dS7AFpUzt1bmPwcFRKcDZwUUE?usp=sharing>. We hope the diversity of DRID-20 will offer unparalleled opportunities to researchers in the multi-instance learning, transfer learning, image dataset construction and other related fields.

This paper is an extended version of [41]. The extensions include: Taking both bag level and instance level noisy images into account in the process of image selection instead of only instance level noisy images, we use a combination of bag level and instance level selection mechanisms and achieve better results; comparing the image classification ability and dataset diversity of our dataset DRID-20 with STL-10, CIFAR-10 and ImageNet; and increasing the number of categories in the dataset from 10 to 20, so that our dataset DRID-20 covers all categories in the PASCAL VOC 2007 dataset.

The rest of the paper is organized as follows: In Section 2, a brief discussion of related works is given. The proposed algorithm including query expanding, noisy expansion filtering and noisy image filtering is described in Section 3. We evaluate the performance of the proposed algorithm against several other methods in Section 4. Lastly, the conclusion and future work are offered in Section 5.

II. RELATED WORKS

Given the importance of labelled image datasets in the area of high-level image understanding, many efforts have been directed toward image dataset construction. In general, these efforts can be divided into three principal categories: manual methods, semi-automatic methods and automatic methods.

A. Manual and Semi-automatic Methods

In the early years, manual labelling was the most important way to construct image datasets. (e.g., STL-10 [29], CIFAR-10 [15], PASCAL VOC 2007 [10], ImageNet [2] and Caltech-101 [32]). The process of constructing these datasets mainly consists of submitting keywords to an image search engine to download candidate images, then cleaning these candidate images by manual annotation. This method has high accuracy but is labor intensive.

To reduce the cost of manual labelling, a large number of works have focused on active learning (a special case of semi-supervised method)[33][36][37]. [33] randomly labelled some seed images to learn visual classifiers. The learned visual classifiers were then implemented to conduct image classification on unlabelled images and find low confidence images for manual labelling. Here low confidence images are those whose probability is classified into positive and negative close to 0.5. The process is iterated until sufficient classification accuracy is achieved. [36] presented an active learning framework to simultaneously learn contextual models for scene understanding tasks (multi-class classification). [37] presented an approach for on-line learning of object detectors, in which the system automatically refines its models by actively requesting crowd-sourced annotations on images crawled from the web. However, both manual labelling and active learning require pre-existing annotations, which often results in one of the most significant limitations to developing a large scale image dataset.

B. Automatic Methods

Automatic methods have attracted more and more attention to decrease the cost of manual annotation [23], [12], [28], [17]. [23] adopted text information to re-rank images retrieved from a web search and used these top-ranked images to learn visual models to re-rank images once again. [17] leveraged the first few images returned from an image search engine to train the image classifier, classifying images as positive or negative. When the image is classified as a positive sample, the classifier uses incremental learning to refine its model. With the increase in the number of positive images accepted by the classifier, the trained classifier will reach a robust level for this query. [12] proposed the use of a clustering based method to filter “group” noisy images and a propagation-based method to filter individual noisy images. The advantage of these methods is that the need for manual intervention is eliminated. However, for methods [23], [12], [17], the domain adaptation ability is limited by the restriction of the initial candidate images and the iterative mechanism in the process of image selection. To obtain high diversity candidate images, [28] proposed the use of multiple query expansions instead of a single query in the

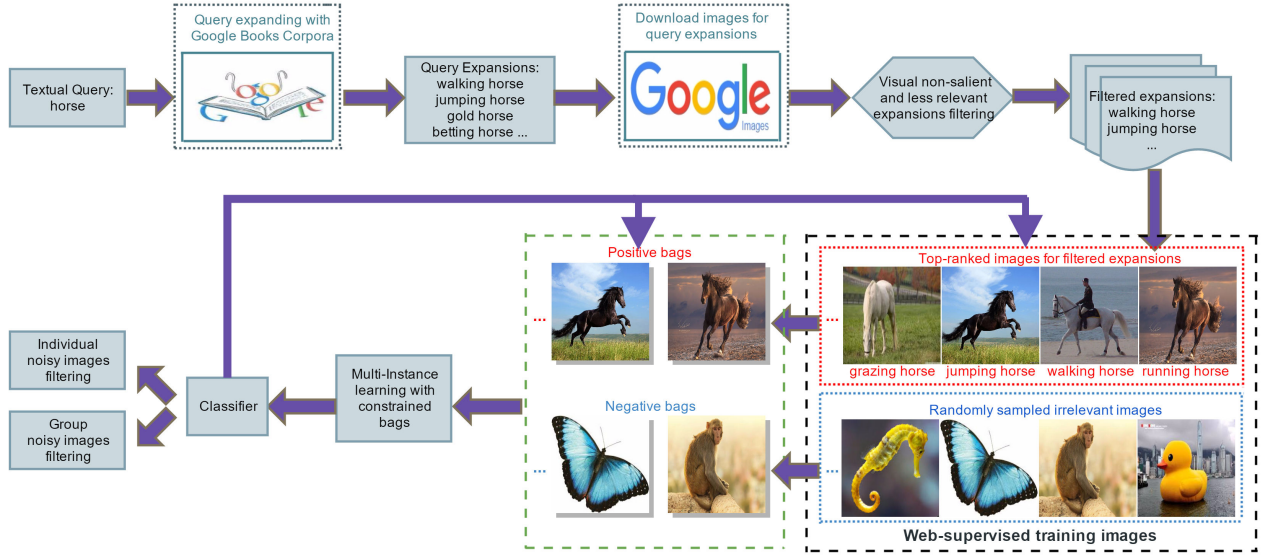


Fig. 2: Flowchart of the proposed approach.

process of collecting the initial candidate images, then using an iterative mechanism to filter noisy images. The automatic works discussed here mainly focus on accuracy and scale in the process of image dataset construction, which often results in poor performance on domain adaptation.

C. Other Related Works

There are many works related to the generation of query expansions and noisy image filtering, though they are not aimed at image dataset construction. Since most image search engines restrict the number of images returned for each query, WordNet [20] and ConceptNet [39] are often used to obtain synonyms to overcome the download restriction of these search engines. The advantage of WordNet and ConceptNet is that synonyms are usually relevant to the given query and almost do not need to be purified. The disadvantage of WordNet and ConceptNet is that both of them are usually not comprehensive enough for query expanding. Worse, the images returned from an image search engine using synonyms tend to experience the homogenization problem, which results in poor performance on domain adaptation. Recent works [28], [9] have proposed the use of Google Books Ngram Corpus (GBNC) [19] to expand query to a set of query expansions. The Google Books Ngrams Corpus covers almost all related queries at the text level. It is much more general and richer than WordNet and ConceptNet. The disadvantage of using GBNC for query expanding is that it may also generate noisy query expansions. Recently, word embedding [8], [40] provides a learning-based method for computing the word-word similarity distance which can be used to filter noisy query expansions. In this paper, we use GBNC to expand the query to a set of query expansions, and then take both word-word and visual-visual similarity distance to filter noisy query expansions.

To efficiently ease the dataset bias problem, several authors have developed domain adaptation approaches for vision tasks. [31] clustered relevant images using both textual and visual features. By treating each cluster as a “bag” and the images in the bag as “instances”, the authors formulated this problem as a multi-instance learning problem (MIL) which learns a target decision function for image re-ranking. However, the yield is limited by the restriction placed on the initial candidate images obtained from the Internet using a single query. In this paper, we focus on the MIL method, as it retains images from different data distributions while filtering out noisy images.

It can be anticipated that there will be more visual patterns (responding to different query expansions) in our work to represent the given query. In addition, MIL methods are applied to filter group and individual noisy images to retain images from different distributions. In return, the constructed image dataset could achieve a better domain adaptation ability than traditional datasets constructed by a single query and an iterative mechanism.

III. PROPOSED APPROACH

We seek to construct a domain robust image dataset which can be well generalized to unseen target domains. As shown in Fig. 2, we propose three major steps for our web-supervised image dataset construction framework: query expanding, noisy expansion filtering and noisy image filtering. A set of semantically rich expansions are obtained by searching in the GBNC [19], from which the visually non-salient and less relevant expansions are filtered by exploiting both word-word and visual-visual similarity. After obtaining the candidate images by retrieving unfiltered expansions with the image search engine, we treat each unfiltered expansion as a “bag” and the images in each bag as “instances”. We then formulate this task as an MIL problem with constrained positive bags. Using this approach, images from different data distributions will be kept

while noisy images will be filtered out, and a domain robust image dataset will be constructed.

A. Query Expanding

Image datasets constructed by existing methods tend to have high accuracy but usually have weak domain adaptation ability [21], [25], [41]. To construct a domain-robust image dataset, we expand the query (e.g., “horse”) to a set of query expansions (e.g., “jumping horse, walking horse, roaring horse”) and then use these different query expansions (corresponding images) to reflect the different “visual patterns” of the query. We use GBNC to discover query expansions for the given query with Parts-Of-Speech (POS), specifically with NOUN, VERB, ADJECTIVE and ADVERB. Our motivation is to identify all related query expansions. GBNC is much more general and richer than WordNet [20] and ConceptNet [39]. Using GBNC can help us to find all the expansions ever published for any possible query.

B. Filtering Noisy Query Expansions

Through query expanding, we obtain a comprehensive semantic description for the given query. However, query expanding not only brings all the useful query expansions, but also some noisy query expansions. These noisy query expansions can be roughly divided into two types: (1) visual non-salient (e.g., “betting horse”) and (2) less relevant (e.g., “sea horse”). Using these noisy query expansions to retrieve images will have a negative impact on dataset accuracy and robustness.

1) *Visually non-salient expansions filtering*: From the visual perspective, we want to identify visually salient query expansions and eliminate non-salient query expansions in this step. The intuition is that visually salient expansions should exhibit predictable visual patterns, hence we use an image classifier-based filtering method. For each query expansion, we directly download the top N images from the Google image search engine as positive images (based on the fact that the top few images returned from image search engine tend to be positive), then randomly split these images into a training set and validation set $I_i = \{I_i^t, I_i^v\}$. We gather a random pool of negative images and split them into a training set and validation set $\bar{I} = \{\bar{I}^t, \bar{I}^v\}$. We train a linear support vector machine (SVM) classifier C_i with I_i^t and \bar{I}^t using dense histogram of oriented gradients (HOG) features. We then use $\{I_i^v, \bar{I}^v\}$ as validation images to calculate the classification results. We declare a query expansion i to be visually salient if the classification results S_i give a relatively high score.

2) *Less relevant expansions filtering*: From the relevance perspective, we want to identify both semantically and visually relevant expansions for the given query. The intuition is that relevant expansions should have a relatively small semantic and visual distance, therefore we use a combined word-word and visual-visual similarity distance-based filtering method.

Words and phrases acquire meaning from the way they are employed in society. For computers, the equivalent of “society” is “database”, and the equivalent of “use” is “a way to search the database” [8]. Normalized Google Distance

(NGD) constructs a method to extract semantic similarity distance from the World Wide Web (WWW) using Google page counts [8]. For a search term x and search term y , NGD is defined by:

$$\text{NGD}(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log N - \min\{\log f(x), \log f(y)\}} \quad (1)$$

where $f(x)$ denotes the number of pages containing x , $f(x, y)$ denotes the number of pages containing both x and y and N is the total number of web pages searched by Google.

We denote the semantic distance of all query expansions by a graph $G_g = \{N, D\}$ in which each node represents a query expansion and its edge represents the NGD between the two nodes. We set the target query as center (x) and other query expansions have a score (D_{xy}) which corresponds to the distance to the target query. Similarly, we represent the visual distance of the query and expansions by a graph $G_v = \{C, E\}$ in which each node represents a query expansion and each edge represents the visual distance between the query and the expansions. The feature is a 1000 dimensional bag of visual words based on SIFT features. The edge weight E_{xy} corresponds to the Euclidean distance.

The semantic distance and visual distance will be used to construct a new two-dimensional feature $V = [D_{xy}; E_{xy}]$. The problem is to calculate the importance weight w and bias penalty b in decision function $f(x) = w^T x + b$ to determine whether or not the expansion is relevant. There are many methods of obtaining these coefficients w and b . Here we take linear SVM to work around this problem. Although linear SVM is not the prevailing state-of-the-art method for classification, we find our method to be effective in pruning irrelevant query expansions.

Unfiltered expansions are then used to retrieve the top M images from the image search engine to construct the raw image dataset. Regardless of the fact that our method is unable to remove noisy expansions thoroughly in most cases, the raw image dataset constructed by our method still achieves much higher accuracy than directly using the Flickr or Google image data. Besides, the raw image dataset constructed through unfiltered query expansions has much richer visual patterns.

C. Filtering Noisy Images

Although the Google image search engine ranks the returned images, several noisy images are still included. In addition, a few unfiltered noisy expansions will also bring noisy images to the raw image dataset. In general, these noisy images can be divided into two types: group noisy images (caused by unfiltered noisy expansions) and individual noisy images (as a result of the error index of the image search engine). To filter these group and individual noisy images while retaining the images from different distributions, we use MIL methods instead of an iterative mechanism in the process of noisy image filtering.

By treating each unfiltered expansion as a “bag” and the images corresponding to the expansion as “instances”, we formulate a multi-instance learning problem by selecting a subset of bags and a subset of images from each bag to

construct a domain robust image dataset for the given query. Since the precision of images returned from the Google image search engine tends to have relatively high accuracy, we define each positive bag as at least having a portion of δ positive instances which effectively filter group noisy images caused by unfiltered noisy query expansions.

We denote each instance as x_i with its label $y_i \in \{0, 1\}$, where $i=1, \dots, n$. We also denote the label of each bag B_I as $Y_I \in \{0, 1\}$. The transpose of a vector or matrix is represented by superscript $'$ and the element-wise product between two matrices is represented by \odot . We define the identity matrix as \mathbf{I} and $\mathbf{0}, \mathbf{1} \in \mathbb{R}^n$ denote the column vectors of all zeros and ones, respectively. The inequality $\mathbf{u} = [u_1, u_2, \dots, u_n]' \geq \mathbf{0}$ means that $u_i \geq 0$ for $i=1, \dots, n$.

1) *Filtering individual noisy images:* The decision function for filtering individual noisy images is assumed in the form of $f(x) = w'\varphi(x) + b$ and has to be learned from the raw image dataset. We employ the formulation of Lagrangian SVM, in which the square bias penalty b^2 and the square hinge loss for each instance are used in the objective function. The decision function can be learned by minimizing the following structural risk function:

$$\min_{\mathbf{y}, \mathbf{w}, b, \rho, \varepsilon_i} \frac{1}{2} \left(\|\mathbf{w}\|^2 + b^2 + C \sum_{i=1}^n \varepsilon_i^2 \right) - \rho \quad (2)$$

$$\text{s.t. } y_i(w'\varphi(x_i) + b) \geq \rho - \varepsilon_i, i = 1, \dots, n. \quad (3)$$

$$\sum_{i: x_i \in B_I} \frac{y_i + 1}{2} \geq \delta |B_I| \quad \text{for } Y_I = 1, \quad (4)$$

$$y_i = 0 \quad \text{for } Y_I = 0$$

where φ is a mapping function that maps x from the original space into a high dimensional space $\varphi(x)$, $C > 0$ is a regularization parameter and ε_i values are slack variables. The margin separation is defined as $\rho / \|\mathbf{w}\|$. $\mathbf{y} = [y_1, \dots, y_n]'$ means the vector of instance labels, $\lambda = \{\mathbf{y} | y_i \in \{0, 1\}\}$ and \mathbf{y} satisfies constraint (4). By introducing a dual variable α_i for inequality constraint (3) and kernel trick $k_{ij} = \varphi(x_i)'\varphi(x_j)$, we arrive at the optimization problem below:

$$\min_{\mathbf{y} \in \lambda} \max_{\alpha} -\frac{1}{2} \left(\sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j k_{ij} + \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j + \frac{1}{C} \right) \quad (5)$$

where $\alpha_i \geq 0$, $\sum_{i=1}^n \alpha_i = 1$ and $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n]'$. By defining $\mathbf{K} = [k_{ij}]$ as a $n \times n$ kernel matrix, $\nu = \{\alpha | \alpha \geq \mathbf{0}, \alpha' \mathbf{1} = 1\}$ and $\tilde{\mathbf{K}} = \mathbf{K} + \mathbf{1}\mathbf{1}'$ as a $n \times n$ transformed kernel matrix for the augmented feature mapping $\tilde{\varphi}(x) = [\varphi(x)', 1]'$ of kernel $\tilde{k}_{ij} = \tilde{\varphi}(x_i)'\tilde{\varphi}(x_j)$. (5) can be rewritten as follows:

$$\min_{\mathbf{y} \in \lambda} \max_{\alpha \in \nu} -\frac{1}{2} \alpha' (\tilde{\mathbf{K}} \odot \mathbf{y}\mathbf{y}' + \frac{1}{C} \mathbf{I}) \alpha \quad (6)$$

(6) is a mixed integer programming problem with respect to the instance labels $y_i \in \{0, 1\}$. We take the Label-Generating MMC (LG-MMC) algorithm proposed in [18] to solve this mixed integer programming problem. We first consider interchanging the order of $\max_{\alpha \in \nu}$ and $\min_{\mathbf{y} \in \lambda}$ in (6) and obtain:

$$\max_{\alpha \in \nu} \min_{\mathbf{y} \in \lambda} -\frac{1}{2} \alpha' (\tilde{\mathbf{K}} \odot \mathbf{y}\mathbf{y}' + \frac{1}{C} \mathbf{I}) \alpha. \quad (7)$$

Algorithm 1 Cutting-plane algorithm for (10)

- 1: Initialize $y_i = Y_I$ for $x_i \in B_I$ as \mathbf{y}^1 , and set $\zeta = \{\mathbf{y}^1\}$;
 - 2: Use MKL to solve α and \mathbf{u} in (10) with ζ ;
 - 3: Select most violated \mathbf{y}^t with α and set $\zeta = \mathbf{y}^t \cup \zeta$;
 - 4: Repeat step 2 and step 3 until convergence.
-

According to the minmax theorem [14], the optimal objective of (6) is an upper bound of (7). We rewrite (7) as:

$$\max_{\alpha \in \nu} \left\{ \max_{\mathbf{y}^t \in \lambda} -\theta | \theta \geq \frac{1}{2} \alpha' (\tilde{\mathbf{K}} \odot \mathbf{y}^t \mathbf{y}^{t'} + \frac{1}{C} \mathbf{I}) \alpha, \forall \mathbf{y}^t \in \lambda \right\} \quad (8)$$

\mathbf{y}^t is any feasible solution in λ . For the inner optimization sub-problem, let $u_t \geq 0$ be the dual variable for inequality constraint. Its Lagrangian can be obtained as:

$$-\theta + \sum_{t: \mathbf{y}^t \in \lambda} u_t \left(\theta - \frac{1}{2} \alpha' (\tilde{\mathbf{K}} \odot \mathbf{y}^t \mathbf{y}^{t'} + \frac{1}{C} \mathbf{I}) \alpha \right). \quad (9)$$

Setting the derivative of (9) with respect to θ to zero, we have $\sum u_t = 1$. $\mathbf{M} = \{\mathbf{u} | \sum u_t = 1, u_t \geq 0\}$ is denoted as the domain of \mathbf{u} , where \mathbf{u} is the vector of u_t . The inner optimization sub-problem is replaced by its dual and (8) can be rewritten as:

$$\max_{\alpha \in \nu} \min_{\mathbf{u} \in \mathbf{M}} -\frac{1}{2} \alpha' \left(\sum_{t: \mathbf{y}^t \in \lambda} u_t \tilde{\mathbf{K}} \odot \mathbf{y}\mathbf{y}' + \frac{1}{C} \mathbf{I} \right) \alpha$$

or

$$\min_{\mathbf{u} \in \mathbf{M}} \max_{\alpha \in \nu} -\frac{1}{2} \alpha' \left(\sum_{t: \mathbf{y}^t \in \lambda} u_t \tilde{\mathbf{K}} \odot \mathbf{y}\mathbf{y}' + \frac{1}{C} \mathbf{I} \right) \alpha. \quad (10)$$

Here, we can interchange the order of $\max_{\alpha \in \nu}$ and $\min_{\mathbf{u} \in \mathbf{M}}$ because the objective function is concave in α and convex in \mathbf{u} . Additionally, (10) can be regarded as a multiple kernel learning (MKL) problem [6], and the target kernel matrix is a convex combination of base kernel matrices $\{\tilde{\mathbf{K}} \odot \mathbf{y}_t \mathbf{y}_t'\}$. Although λ is finite and (10) is an MKL problem, we can not directly use existing MKL techniques like [22] to solve this problem. The reason is that the exponential number of possible labellings $\mathbf{y}_t \in \lambda$ and the fact that the base kernels are also exponential in size make direct MKL computations intractable.

Fortunately, not all the constraints in (8) are active at optimality, thus we can employ a cutting-plane algorithm [13] to find a subset $\zeta \in \lambda$ of the constraints that can well approximate the original optimization problem. The detailed solutions of the cutting-plane algorithm for (10) are described in Algorithm 1. Finding the most violated constraint \mathbf{y}^t is the most challenging aspect of the cutting-plane algorithm. According to (5), the most violated \mathbf{y}^t is equivalent to the following optimization problem:

$$\max_{\mathbf{y} \in \lambda} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j k_{ij}. \quad (11)$$

We solve this integer optimization problem by enumerating all possible candidates of \mathbf{y}^t . Here we only enumerate the possible labelling candidates of the instances in positive bags as all instances in the negative bags are assumed to be negative

in our paper. Lastly, we can derive the decision function from the raw image dataset for the given query as:

$$f(x) = \sum_{i: \alpha_i \neq 0} \alpha_i \tilde{y}_i \tilde{k}(x, x_i) \quad (12)$$

where $\tilde{y}_i = \sum_{t: y^t \in \lambda} u_t y_i^t$ and $\tilde{k}(x, x_i) = k(x, x_i) + 1$. The decision function will be used to filter individual noisy images in each bag which correspond to unfiltered query expansions.

2) *Filtering group noisy images:* To filter group noisy images (caused by unfiltered noisy expansions), we represent bag B_I with the compound feature $\phi_{f,k}$ of its first k positive instances:

$$\phi_{f,k}(B_I) = \frac{1}{k} \sum_{x_i \in \Psi_{f,k}^*(B_I)} x_i \quad (13)$$

with

$$\Psi_{f,k}^*(B_I) = \arg\max_{\Psi \subseteq B_I, |\Psi|=k} \sum_{x_i \in \Psi} f(x_i). \quad (14)$$

We refer to the instances in $\Psi_{f,k}^*(B_I)$ as the first k instances of B_I according to classifier f (see Equation 12). The closer the images in B_I are to the bag centre, the higher is the probability that these images will be relevant to the bag. The assignment of relatively heavier weights to images which are a short distance from the bag centre will increase the accuracy of classifying bag B_I as positive or negative, then increase the efficiency of filtering noisy group images. Following [42], we assume $\xi_i = [1 + \exp(\alpha \log d(x_i) + \beta)]^{-1}$ to be a weighting function, $d(x_i)$ represents the Euclidean distance of images x_i from the bag centre, $\alpha \in \mathbb{R}_{++}$ and β are scaling and offset parameters which can be determined by cross-validation. The representation of (13) for bag B_I can be generalized to a weighted compound feature:

$$\phi_{f,k}(B_I) = \phi(X, h^*) = \frac{Xh^*}{\xi^T h^*} \quad (15)$$

with

$$h^* = \arg\max_{h \in H} f\left(\frac{Xh}{\xi^T h}\right), \quad \text{s.t.} \quad \sum_i h_i = k \quad (16)$$

where $X = [x_1, x_2, x_3, \dots, x_i] \in \mathbb{R}^{D \times i}$ is a matrix whose columns are the instances of bag B_I , $\xi = [\xi_1, \xi_2, \xi_3, \dots, \xi_i]^T \in \mathbb{R}_{++}^i$ are the vectors of weights, and $h^* \in H = \{0, 1\}^i \setminus \{0\} \mid (\sum_i h_i = k)$ is an indicator function for the first k positive instances of bag B_I .

Then classifying rule of bag B_I to be positive or negative is:

$$f_\omega(X) = \max_{h \in H} \omega^T \phi(X, h), \quad \sum_i h_i = k \quad (17)$$

where $\omega \in \mathbb{R}^D$ is the vector of classifying coefficients, $\phi(X, h) \in \mathbb{R}^D$ is the feature vector of (15), h is a vector of latent variables and H is the hypothesis space $\{0, 1\}^i \setminus \{0\}$. The learning problem is to determine the parameter vector ω . Given a training set $\tau = \{B_I, Y_I\}_{I=1}^n$, this is a latent SVM learning problem:

$$\min_{\omega} \frac{1}{2} \|\omega\|^2 + C \sum_{I=1}^n \max(0, 1 - Y_I f_\omega(X_{B_I})). \quad (18)$$

Algorithm 2 Concave-convex procedure for solving (21)

- 1: Initialize ω with SVM by setting $h = \mathbf{1} \in \mathbb{R}^i$;
 - 2: Compute a convex upper bound using the current model for the second term of (21);
 - 3: Minimize this upper bound by solving a structural SVM problem via the proximal bundle method [38];
 - 4: Repeat step 2 and step 3 until convergence.
-

Before solving (18), we first solve the classifying rule of (17). It is necessary to solve the below following problem:

$$\max_{h \in H} \frac{\omega^T Xh}{\xi^T h}, \quad \text{s.t.}, \quad \sum_i h_i = k. \quad (19)$$

This is an integer linear-fractional programming problem. Since $\xi \in \mathbb{R}_{++}^i$, (19) is identical to the relaxed problem:

$$\max_{h \in \beta^i} \frac{\omega^T Xh}{\xi^T h}, \quad \text{s.t.}, \quad \sum_i h_i = k \quad (20)$$

where $\beta^i = [0, 1]^i$ is a unit box in \mathbb{R}^i . (20) is a linear-fractional programming problem and can be reduced to a linear programming problem of $i + 1$ variables and $i + 2$ constraints [43].

In this work, we take the concave-convex procedure (CCCP) [44] to solve (18). We rewrite the objective of (18) as two convex functions:

$$\min_{\omega} \left[\frac{1}{2} \|\omega\|^2 + C \sum_{I \in D_n} \max(0, 1 + f_\omega(X_{B_I})) + C \sum_{I \in D_p} \max(f_\omega(X_{B_I}), 1) \right] - \left[C \sum_{I \in D_p} f_\omega(X_{B_I}) \right] \quad (21)$$

where D_p and D_n are positive and negative training sets respectively. The detailed solutions of the CCCP algorithm for (21) are described in Algorithm 2. Lastly, we obtain the bag classifying rule as (17) to filter group noisy images which correspond to unfiltered noisy query expansions.

In summary, the existing automatic methods reduce the cost of manual annotation by leveraging the generalization ability of machine learning models. However, this generalization ability is affected by both the quality of the initial candidate images and the capability of models to retain images from different distributions. Previous works largely focus on accuracy and scale, and most use an iterative mechanism for the image selection process which often results in the dataset bias problem. To the best of our knowledge, we are the first to propose the automatic construction of a domain-robust image dataset. We achieve the domain adaptation ability of our dataset by maximizing both the initial candidate images and the final selected images from different data distributions.

IV. EXPERIMENTS

Since the datasets for existing dataset construction methods [12], [23], [24], [17] have not been released, we are unable to directly compare our dataset with their extracted datasets.

We therefore systematically compare the image classification ability, cross-dataset generalization ability and dataset diversity of our dataset with three publicly available datasets STL-10, CIFAR-10 and ImageNet. The motivation is to verify that a domain-robust image dataset has a better image classification ability on third-party datasets, and to confirm that a domain-robust image dataset has better cross-dataset generalization ability and dataset diversity. We also report the object detection ability of our dataset and compare our method with four baseline methods [9], [24], [27], [35].

A. Image Dataset DRID-20 Construction

To facilitate comparison with datasets STL-10, CIFAR-10 and ImageNet, we choose common categories in these datasets: airplane/aeroplane, bird, cat, dog, horse to construct our dataset. We also select 15 other categories in PASCAL VOC 2007 to construct our dataset, since most of the existing weakly supervised and web-supervised learning methods are tested on the PASCAL VOC 2007 dataset. Overall, we use the proposed method in this paper to build our dataset, DRID-20, which consists of all 20 categories in the PASCAL VOC 2007 dataset.

For each given query (e.g., “horse”) in our experiments, we first expand the given query to a set of query expansions with POS. To filter visual non-salient expansions, we retrieve the top $N = 100$ images from the image search engine as positive images (in spite of the fact that noisy images might be included). Set the training set and validation set $I_i = \{I_i^t = 75, I_i^v = 25\}$, $\bar{I} = \{\bar{I}^t = 25, \bar{I}^v = 25\}$. Through experiments, we declare a query expansion i to be visually salient if the classification results ($S_i \geq 0.7$) return a relatively high score. We have released the query expansions for twenty categories and corresponding images (original image URL on website: <https://drive.google.com/drive/folders/0B7dS7AFpUzt1bnpwvFRKcdZwUUE?usp=sharing>).

To filter less relevant expansions, we select n_+ positive training samples from these expansions that have a small semantic or visual distance from these expansions. We calculate the semantic distance and visual distance between different queries (e.g., “horse” and “cow”) and obtain the n_- negative training samples. We do not select the n_- negative training samples from expansions which have a large semantic or visual distance because these expansions have a higher probability of being positive than other different query expansions. Here, we set $n = 1000$ and train a classifier based on linear SVM to filter less relevant expansions.

The first $M = 100$ (for category “plant” expansions, $M = 350$) images are retrieved from the Google image search engine for each unfiltered query expansion to construct the raw image dataset. We treat unfiltered query expansions as positive bags and images in bags as instances. We define each positive bag as having at least a portion of $\delta = 0.7$ positive instances. Negative bags can be obtained by randomly sampling a few irrelevant images that are not associated with the given query. MIL methods are applied to learn the decision function (12), and the individual noisy images in each bag are filtered. The decision function of (12) is also used to select the most k

TABLE I: The number of images for each category in our experiments

Dataset \ Category	airplane	bird	cat	dog	horse
STL-10	1300	1300	1300	1300	1300
CIFAR-10	6000	6000	6000	6000	6000
PASCAL VOC	238	330	337	421	287
ImageNet	1434	2126	1083	1603	1402
DRID-20	1000	1000	1000	1000	1000

positive instances in each bag, representing this bag for group noisy image filtering. The value of k for different categories may be different. In general, categories which have larger query expansions tend to select a smaller value. There are multiple methods for learning the weighting function (e.g., logistic regression or cross-validation), here we follow [42] and use cross-validation to learn the weighting function. We label 10 datasets, each containing 100 positive bags and 100 negative bags. The positive bags and negative bags each have 50 images. Labelling only needs to be carried out once to learn the weighting function and weighted bag classification rule (17). The learned weighted bag classification rule (17) will be used to filter noisy bags (corresponding to group noisy images). For better comparison with other datasets, we evenly select positive images from positive bags to construct the dataset DRID-20. Each category in DRID-20 has 1000 images, and this dataset has been released publicly on website.

B. General experimental set-up

For dataset image classification ability, cross-dataset generalization ability and dataset diversity comparison, we select five common categories in STL-10, CIFAR-10, PASCAL VOC 2007, ImageNet and DRID-20. For object detection ability comparison, we use all 20 categories in the DRID-20 dataset and PASCAL VOC 2007.

1) *General setting for image classification, cross-dataset generalization and dataset diversity:* STL-10 has ten categories, and each category of which contains 500 training images and 800 test images. All of the images in STL-10 are colour 96×96 pixels. We use all the training images and test images in STL-10 to represent the dataset. The CIFAR-10 dataset consists of 60000 32×32 colour images in 10 categories, with 6000 images per category, of which 5000 are training images and 1000 are test images. Similarly, we use all 6000 images in CIFAR-10 to represent the dataset. ImageNet is an image dataset organized according to the WordNet [20] hierarchy. It provides on average 1000 images to illustrate each category. We use all the images in ImageNet for each category to represent the ImageNet dataset. PASCAL VOC 2007 is a benchmark dataset in image classification and object detection which provides the vision and machine learning communities with a standard dataset of images and evaluation procedures. PASCAL VOC 2007 contains 20 categories, each of which contains training/validation data and test data. For image classification ability, cross-dataset generalization ability and dataset diversity comparison, we utilize the training/validation data to represent the PASCAL VOC 2007 dataset. Our dataset DRID-20 is constructed according

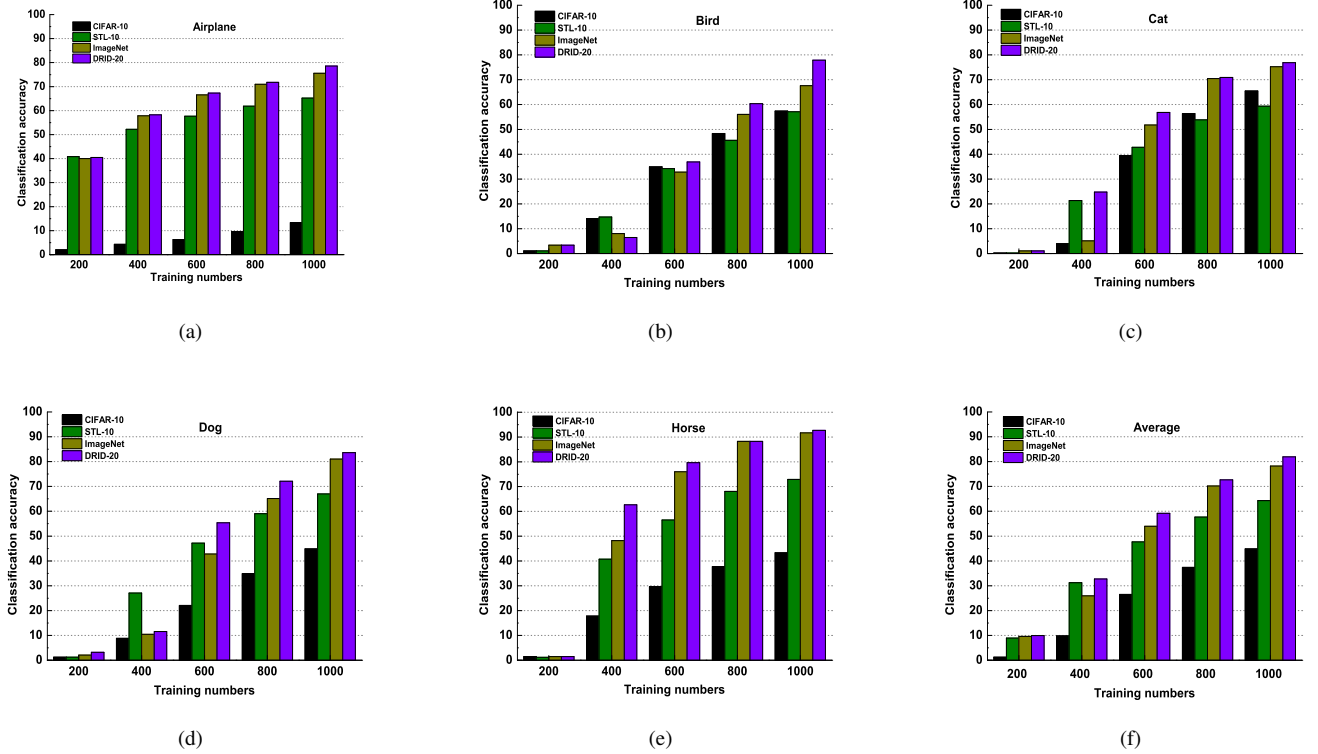


Fig. 3: Image classification ability of CIFAR-10, STL-10, ImageNet and DRID-20 on PASCAL VOC 2007 dataset: (a) airplane, (b) bird, (c) cat, (d) dog, (e) horse and (f) average.

to the categories in PASCAL VOC 2007 and has 1000 images in each category. To evaluate the image classification ability, cross-dataset generalization ability and dataset diversity, we resize all the images in STL-10, ImageNet, PASCAL VOC 2007 and our DRID-20 to 32×32 . For all datasets, we extract the same Histogram of Oriented Gradient (HOG) feature and train one-versus-all classifiers. The detailed number of images in each category for above mentioned experiments is shown in Table 1.

2) *General setting for object detection*: The idea of training detection models without bounding boxes has received renewed attention due to the success of the DPM [27] detector. To compare the object detection ability of our method with four other baseline methods [9], [27], [24], [35], we select PASCAL VOC 2007 as the test data, because recent state-of-the-art weakly supervised and web-supervised methods have been evaluated on this dataset.

For each query expansion, we train a separate DPM to constrain the visual variance. We resize images to a maximum of 500 pixels and ignore images with extreme aspect ratios (aspect ratio > 2.5 or < 0.4). To avoid getting stuck to the image boundary during the latent re-clustering step, we initialize our bounding box to a sub-image within the image that ignores the image boundaries. Following [27], we also initialize components using the aspect-ratio heuristic. Some of the components across different query expansion detectors ultimately learn the same visual pattern. For example, the images corresponding to the query expansion “walking horse”

are similar to the images corresponding to “standing horse”. In order to select a representative subset of the components and merge similar components, we represent the space of all query expansions components by a graph $G = \{C, E\}$, in which each node represents a component and each edge represents the visual similarity between them. The score d_i for each node corresponds to the average precision. The weight on each edge $e_{i,j}$ is obtained by running the j th component detector on the i th component set. We solve for the same objective function proposed in [9] to select the representative components $S \subseteq V$:

$$\max_S \sum_{i \in V} d_i \cdot \vartheta(i, S) \quad (22)$$

where ϑ is a soft coverage function that implicitly pushes for diversity:

$$\vartheta(i, S) = \begin{cases} 1 & i \in S \\ 1 - \prod_{j \in S} (1 - e_{i,j}) & i \notin S. \end{cases} \quad (23)$$

After the representative subset of components has been obtained, we augment them with method as described in [27] and subsequently merge all the components to produce the final detector.

C. Performance Evaluation of Image Classification Ability

We choose PASCAL VOC 2007 as the third-party test data for comparing the image classification ability of our dataset DRID-20 with STL-10, CIFAR-10 and ImageNet. For this experiment, we select five categories that are common to all

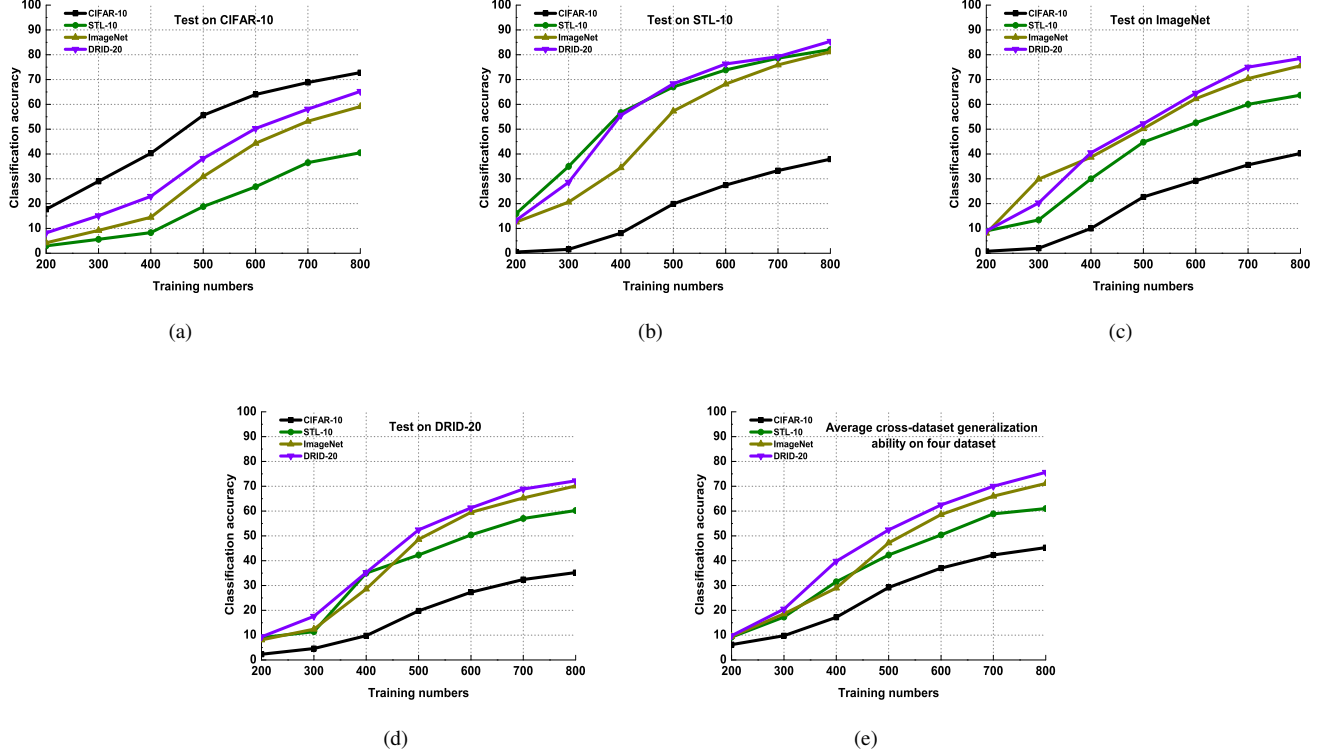


Fig. 4: Cross-dataset generalization performance of classifier learned from different datasets and then tested on: (a) CIFAR-10, (b) STL-10, (c) ImageNet, (d) DRID-20, (e) Average.

these datasets: airplane/aeroplane, bird, cat, dog and horse. We randomly select training images from various datasets for our choice of positive training images. We choose the same 1000 negative training images for all datasets. For details, we sequentially select [200,400,600,800,1000] training images from CIFAR-10, STL-10, ImageNet and DRID-20 as the positive training images, and use 1000 fixed negative training images to learn the image classifiers. We then test the performance of these classifiers on the corresponding categories of the PASCAL VOC 2007 dataset. We repeat the above experiment 10 times and use the average performance of image classifiers as the final performance for each dataset. The image classification ability of all datasets for each category and the entire dataset is shown in Figure 3.

We make the following observations from Fig. 3:

(1) It is interesting to observe that the category “airplane” has a relatively higher classification accuracy than the categories “bird”, “cat”, “dog” and “horse” with a small amount of training data [200,400]. A possible explanation is that the scenes and visual patterns of “airplane” are relatively simpler than the categories “bird”, “cat”, “dog” and “horse”. Even with a small amount of training data, there are still a large number of positive patterns in both auxiliary and target domains. That is to say, the samples are densely distributed in the feature space, and the distribution of the two domains overlap much more easily. On the other hand, the positive samples from both domains for the categories “bird”, “cat”, “dog” and “horse” are distributed sparsely in the feature space. It is likely that

there will be less overlap of the data distributions of the two domains.

(2) CIFAR-10 exhibits much worse performance on image classification than STL-10, ImageNet and DRID-20 according to the accuracy over all five common categories, which demonstrates that the SVM classifier learned with training data from the auxiliary domain performs poorly on the target domain. The explanation is perhaps that the data distributions of CIFAR-10 are quite different from those of the PASCAL VOC 2007 dataset. The CIFAR-10 dataset has a more serious dataset bias problem than STL-10, ImageNet and DRID-20.

(3) We also observe that ImageNet is slightly worse than DRID-20 in each individual category and in the whole dataset, possibly because the distribution of samples from ImageNet is relatively rich. ImageNet is constructed with the goal that objects in images should have variable appearance, positions, view points, and poses, as well as background clutter and occlusions.

(4) DRID-20 outperforms CIFAR-10, STL-10 and ImageNet in terms of average accuracy in five common categories, which demonstrates the domain robustness of DRID-20. The explanation is that DRID-20 constructed by multiple query expansions and MIL selecting mechanisms has much more visual patterns than CIFAR-10, STL-10 and ImageNet when given the same number of training samples. In other words, DRID-20 has much richer feature distribution and is more easily overlapped with unknown target domains.

We also report the hardware configuration of our experi-

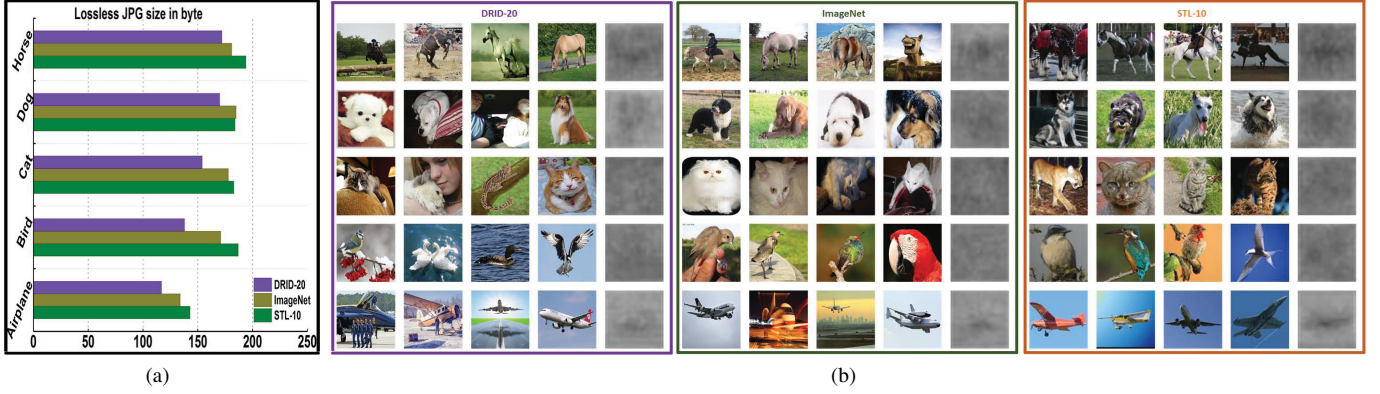


Fig. 5: (a) Comparison of the lossless JPG file sizes of average images for five different categories in DRID-20, ImageNet and STL-10. (b) Example images from DRID-20, ImageNet, STL-10 and average images for each category indicated by (a).

ment. We use HP desktop PCs (3.2GHz CPU with 8 Gbyte RAM) for the image collection. All the data processing and experiments are performed on an Acer workstation (3.5GHz CPU, 16 Gbyte RAM and 4 Gbyte VRAM) with LIBSVM [7].

D. Performance Evaluation of Cross-dataset Generalization Ability

Cross-dataset generalization ability measures the performance of classifiers learned from one dataset and tested on other datasets. It indicates the domain robustness of dataset [25], [28], [41]. Here we compare the cross-dataset generalization ability of our dataset DRID-20 with three publicly available dataset CIFAR-10, STL-10 and ImageNet. We choose the same five categories (horse, bird, airplane, cat and dog) included in all four datasets to verify their cross-dataset generalization ability.

We randomly select 200 images per category from each dataset as the test data. For the choice of training data, we sequentially select [200,300,400,500,600,700,800] images per category from various datasets as the positive training images and use 1000 fixed negative training images to learn the image classifiers. The training images in each category are selected randomly and the training images and test images have no duplicates. The average classification accuracy for five categories (horse, bird, airplane, cat and dog) represents the cross-dataset generalization ability of one dataset on another dataset. When training the image classification model, we set the same options for four datasets; we set the type of SVM as C-SVC, the type of kernel as a radial basis function and all other options as the default LIBSVM options. The cross-dataset performance of the four datasets and their average performance is shown in Figure 4.

By observing Figure 4, we draw the following conclusions:

(1) In three of four datasets, the best classification performance with the increase in the number of training images is achieved by DRID-20. When tested on STL-10, ImageNet and DRID-20, it can be seen that the generalization ability of STL-10, ImageNet and our dataset DRID-20 is very close, but DRID-20 performs slightly better than STL-10 and ImageNet.

In addition, DRID-20 outperforms CIFAR-10, STL-10 and ImageNet in terms of average cross-dataset performance on four datasets, which demonstrates the domain robustness of DRID-20. A possible explanation is that our DRID-20 dataset, being constructed by multiple query expansions, has much more visual patterns or feature distributions than STL-10 or CIFAR-10, which just use only one query for candidate image collection. At the same time, MIL selection mechanisms maximize the retention of useful visual patterns to represent the DRID-20 dataset.

(2) CIFAR-10 shows poor performance on cross-dataset generalization except on its own dataset. The explanation is that the data distributions of its auxiliary domain and target domain are quite strongly related, making it difficult for other datasets to exceed its performance when tested on CIFAR-10. All images in CIFAR-10 are cut to 32×32 and objects in these images are located in the middle of the image. Besides, these images contain relatively small other objects or scenes. Images in STL-10 are 96×96 and are full size in ImageNet and DRID-20. These images not only contain target objects, but also include a large number of other scenarios or objects. Based on these conditions, although CIFAR-10 has a better performance on its own domain, it still has a serious dataset bias problem which coincides with its average cross-dataset generalization performance.

E. Performance Evaluation of Dataset Diversity

Following [2], [33], we compute the average image of each category and measure the lossless JPG file size, which reflects the amount of information in an image. The basic idea is that a diverse image dataset will result in a blurrier average image, whereas an image dataset with little diversity will result in a more structured, sharper average image. Therefore, we expect the average image of a more diverse image dataset to have a smaller JPG file size.

We resize all images in STL-10, ImageNet and DRID-20 to 32×32 grey images, and create average images for each category from 100 randomly sampled images. Fig. 5 compares the image diversity of five common categories in DRID-20, ImageNet, and STL-10, and shows example images

TABLE II: Object detection results (A.P.) on PASCAL VOC 2007 (test).

Method	[24]	[35]	[9]	Our	[27]
Supervision	weak	weak	web	web	full
airplane	13.4	17.4	14.0	15.5	33.2
bike	44.0	-	36.2	40.6	59.0
bird	3.1	9.3	12.5	16.1	10.3
boat	3.1	9.2	10.3	9.69	15.7
bottle	0.0	-	9.2	13.7	26.6
bus	31.2	-	35.0	42.0	52.0
car	43.9	35.7	35.9	37.9	53.7
cat	7.1	9.4	8.4	9.8	22.5
chair	0.1	-	10.0	9.6	20.2
cow	9.3	9.7	17.5	18.4	24.3
table	9.9	-	6.5	10.6	26.9
dog	1.5	3.3	12.9	11.6	12.6
horse	29.4	16.2	30.6	36.1	56.5
motorcycle	38.3	27.3	27.5	36.9	48.5
person	4.6	-	6.0	7.9	43.3
plant	0.1	-	1.5	1.3	13.4
sheep	0.4	-	18.8	20.4	20.9
sofa	3.8	-	10.3	10.8	35.9
train	34.2	15.0	23.5	27.6	45.2
tv/monitor	0.0	-	16.4	18.4	42.1
average	13.87	15.25	17.15	19.74	33.14

and average images in these datasets. By observing Fig. 5(a), it can be seen that the average image of DRID-20 is blurred and it is difficult to recognize the object, while the average image of ImageNet and STL-10 is relatively more structured and sharper. DRID-20 has a slightly smaller JPG file size than ImageNet and STL-10. This phenomenon is universal for all five categories.

DRID-20 is constructed with the goal that images in this dataset should exhibit domain robustness and be able to effectively alleviate the dataset bias problem. To achieve domain robustness, we not only consider the source of the candidate images, but also retain the images from different distributions. We can see from the above experiments that, with a certain number of samples, DRID-20 has much more effective visual patterns and feature distributions than the CIFAR-10, STL-10 and ImageNet datasets, and thus has better domain adaptation ability.

F. Performance Evaluation of Object Detection Ability

We report the performance of object detection on the PASCAL VOC 2007 test set. Table 2 shows the results of our proposed method and compares it to the state-of-the-art weakly supervised and web-supervised methods [9], [24], [35]. Methods [24] and [35] have state-of-the-art performance for weakly supervised object detection. [35] is trained on manually selected videos without bounding boxes and shows results on 10 out of 20 categories. [24] uses weak human supervision (VOC data with image-level labels for training) and initialization from objectness [30]. [9] takes web supervision and then trains a mixture DPM detector for the object. [27] is a fully supervised object detection method and it is a possible upper bound for weakly supervised and web-supervised approaches.

Compared to [24], [35] which uses weak supervision, and [27] which uses full supervision, the training set of our proposed approach and [9] do not need to be labelled manually. Nonetheless, the results of our proposed approach and [9]

surpass the previous best results of weakly supervised object detection methods [24], [35]. A possible explanation is perhaps that both our approach and that of [9] use multiple query expansions for candidate image collection, and the training data collected by our approach and [9] are richer and contain more effective visual patterns. In most cases, our method surpasses the results in [9], which also uses web supervision and multiple query expansion for candidate image collection. The explanation for this is that we use different mechanisms for noisy images removal. Compared to [9] which takes iterative mechanisms in the process of noisy image filtering, our approach applies an MIL method for removing noisy images. This maximizes the ability to retain images from different data distributions while filtering out the noisy images.

By using the same feature and training strategies, our approach achieves the best performance compared to weakly supervised and web-supervised method [24], [35], [9]. The main reason for this is that our training data are generated from multiple query expansions and MIL filtering mechanisms, and thus contain much richer and more accurate visual descriptions for these categories. In other words, our approach discovers many more useful linkages to visual patterns for the given category.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented a new framework for domain-robust image dataset construction with web images. Three successive modules were employed in the framework including query expanding, noisy expansion filtering and noisy image filtering. To verify the effectiveness of our proposed method, we constructed an image dataset DRID-20. Extensive experiments show that our dataset has better domain adaptation ability than the traditional manual-labelled datasets STL-10, CIFAR-10 and ImageNet. In addition, our data was successfully applied to help improve object detection on PASCAL VOC 2007, and the results demonstrated the superiority of our method to several weakly supervised and web-supervised state-of-the-art methods. We have publicly released the DRID-20 dataset to facilitate the research in this field.

Although good results were obtained, there is still room to improve the proposed dataset construction framework. For example, we can potentially use more sophisticated approaches to purify noisy query expansions, noisy images and that will be the focus of our future work.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (No. 61473154).

REFERENCES

- [1] Z. Li and J. Tang, "Weakly supervised deep metric learning for community-contributed image retrieval," *IEEE Transactions on Multimedia*, 17(11): 1989–1999, 2015.
- [2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *IEEE International Conference on Computer Vision and Pattern Recognition*, 248–255, 2009.

- [3] L. Zhang, M. Song, Y. Yang, Q. Zhao, C. Zhao, and N. Sebe, "Weakly supervised photo cropping," *IEEE Transactions on Multimedia*, 16(1): 94–107, 2014.
- [4] Y. Tang, X. Wang, E. Dellandrea, and L. Chen, "Weakly supervised learning of deformable part-based models for object detection via region proposals," *IEEE Transactions on Multimedia*, 2016.
- [5] R. Ewerth, K. Ballafkir, M. Muhling, D. Seiler, and B. Freisleben, "Long-term incremental web-supervised learning of visual concepts via random savannas," *IEEE Transactions on Multimedia*, 14(4): 1008–1020, 2012.
- [6] F. Bach, G. Lanckriet, and M. Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," *ACM International Conference on Machine Learning*, 220–228, 2004.
- [7] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.
- [8] R. Cilibrasi and P. Vitanyi, "The google similarity distance," *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.
- [9] S. Divvala, C. Guestrin, "Learning everything about anything: Webly-supervised visual concept learning," *IEEE International Conference on Computer Vision and Pattern Recognition*, 3270–3277, 2014.
- [10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, 88(2):303–338, 2010.
- [11] P. Felzenszwalb, R. Girshick, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.
- [12] X. Hua and J. Li, "Prajna: Towards recognizing whatever you want from images without image labeling," *AAAI International Conference on Artificial Intelligence*, 137–144, 2015.
- [13] J. Kelley, "The cutting-plane method for solving convex programs," *Journal of the Society for Industrial and Applied Mathematics*, 8(4): 703–712, 1960.
- [14] S. Kim and S. Boyd, "A minimax theorem with applications to machine learning, signal processing, and finance," *SIAM Journal on Optimization*, 19(3): 1344–1367, 2008.
- [15] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *CiteSeer*, 2009.
- [16] T. Leung, Y. Song, and J. Zhang, "Handling label noise in video classification via multiple instance learning," *IEEE International Conference on Computer Vision*, 2056–2063, 2011.
- [17] L. Li and L. Fei-Fei, "Optimol: automatic online picture collection via incremental model learning," *International Journal of Computer Vision*, 88(2):147–168, 2010.
- [18] Y. Li, I. Tsang, J. Kwok, and Z. Zhou, "Tighter and convex maximum margin clustering," *International Conference on Artificial Intelligence and Statistics*, 344–351, 2009.
- [19] Y. Lin, J. Michel, E. Aiden, J. Orwant, W. Brockman, and S. Petrov, "Syntactic annotations for the google books ngram corpus," *ACL 2012 System Demonstrations*, 169–174, 2012.
- [20] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, 38(11): 39–41, 1995.
- [21] L. Niu, W. Li, and D. Xu, "Visual recognition by learning from web data: A weakly supervised domain generalization approach," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2774–2783, 2015.
- [22] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "Simplemkl," *Journal of Machine Learning Research*, 9(1): 2491–2521, 2008.
- [23] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 33(4): 754–766, 2011.
- [24] P. Siva and T. Xiang, "Weakly supervised object detector learning with model drift detection," *IEEE International Conference on Computer Vision*, 343–350, 2011.
- [25] A. Torralba and A. Efros, "Unbiased look at dataset bias," *IEEE International Conference on Computer Vision and Pattern Recognition*, 1521–152, 2011.
- [26] S. Vijayanarasimhan and K. Grauman, "Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization," *IEEE International Conference on Computer Vision and Pattern Recognition*, 1–8, 2008.
- [27] P. Felzenszwalb, R. Girshick, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9): 1627–1645, 2010.
- [28] Y. Yao, J. Zhang, F. Shen, X. Hua, J. Xu, and Z. Tang, "Automatic image dataset construction with multiple textual metadata," *IEEE International Conference on Multimedia and Expo*, 1–6, 2016.
- [29] A. Coates, A. Ng, H. Lee, "An analysis of single-layer networks in unsupervised feature learning," *International Conference on Artificial Intelligence and Statistics*, 215–223, 2011.
- [30] B. Alexe, T. Deselaers, V. Ferrari, "Measuring the objectness of image windows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11): 2189–2202, 2012.
- [31] L. Duan, W. Li, I. Tsang, and D. Xu, "Improving web image search by bag-based reranking," *IEEE Transactions on Image Processing*, 20(11): 3280–3290, 2011.
- [32] G. Griffin, A. Holub, P. Perona, "Caltech-256 object category dataset."
- [33] B. Collins, J. Deng, K. Li, L. Fei-Fei, "Towards scalable dataset construction: An active learning approach, in: Computer Vision–ECCV 2008, Springer, 2008, pp. 86–98.
- [34] J. Xiao, J. Hays, K. Ehinger, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," *IEEE International Conference on Computer Vision and Pattern Recognition*, 3485–3492, 2010.
- [35] A. Prest, C. Leistner, J. Civera, and V. Ferrari, "Learning object class detectors from weakly annotated video," *IEEE International Conference on Computer Vision and Pattern Recognition*, 3282–3289, 2012.
- [36] B. Siddiquie, A. Gupta, "Beyond active noun tagging: Modeling contextual interactions for multi-class active learning," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2979–2986, 2010.
- [37] S. Vijayanarasimhan, K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," *International Journal of Computer Vision*, 108(2), 97–114, 2014.
- [38] K. C. Kiwiel, "Proximity control in bundle methods for convex non differentiable minimization," *Mathematical Programming*, 46(1-3), 105–122, 1990.
- [39] R. Speer, C. Havasi, "Conceptnet 5: A large semantic network for relational knowledge," *The Peoples Web Meets NLP*, 161–176, 2013.
- [40] R. Collobert, J. Weston, "A unified architecture for natural language processing: Deep neural networks with multi-task learning," *ACM International Conference on Machine Learning*, 160–167, 2008.
- [41] Y. Yao, X. Hua, F. Shen, J. Zhang, and Z. Tang, "A domain robust approach for image dataset construction," *ACM International Conference on Multimedia*, 212–216, 2016.
- [42] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3), 394–410, 2007.
- [43] S. Boyd and L. Vandenberghe, "Convex optimization," *Cambridge University Press*, 2004.
- [44] A. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural computation*, 15(4): 915–936, 2003.